



Стандарт перевода с использованием МП (МТРЕ)

Редакция: 1.2

Дата редакции: 04.03.2020

Отдел: *Переводов*

Код документа: *RGL-TRA-008*

Оглавление

I.	Термины и определения	3
II.	Введение	4
III.	Применение МП в работе компании «ИнТекст»	4
IV.	Работа переводчика с использованием МП	5
V.	Советы по работе с машинным переводом.....	6
VI.	Оценивание качества машинного перевода	8
VII.	Нештатные ситуации.....	9
VIII.	Условия оплаты	9
IX.	Лист регистрации изменений	Ошибка! Закладка не определена.

I. Термины и определения

Машинный перевод (МП) / Machine Translation (MT) — процесс перевода текстов с одного естественного языка на другой с помощью специальной компьютерной программы. Он достигается с помощью полностью автоматизированных инструментов без участия человека. Машинный перевод ускоряет процесс и часто используется для перевода большого объема текста.

Машинный перевод на основе правил (МПОП) / Rule-Based Machine Translation (RBMT) — машинный перевод, который основывается на грамматике, закономерностях языка и двуязычных словарях. Самое первое поколение технологии МП.

Машинный перевод на основе примеров (МПП) / Example-based Machine Translation (EBMT) — машинный перевод, который основывается на поиске аналогий между текстом на перевод и существующих параллельных текстах. Программа находит схожие случаи перевода в своих архивах и применяет их для нового текста. Второе поколение технологий МП.

Статистический машинный перевод (СМП) / Statistical Machine Translation (SMT) — машинный перевод, который основывается на анализе большого объема параллельных текстов. Такая система подсчитывает варианты перевода отдельных слов и их частоту, после чего переносит наиболее вероятный вариант в перевод. Третье поколение систем МП.

Нейронный машинный перевод (НМТ) / Neural Machine Translation (NMT) — машинный перевод, который основывается на технологии машинного обучения, нейронной сети. Для этой технологии свойственно использование параллельных корпусов текста и необходимость в значительных вычислительных мощностях. Это четвертое поколение технологий МП. Первые нейронные движки вышли на рынок в 2016 году, после чего произошло существенное улучшение качества МП.

Движок машинного перевода (ДМП) / Machine Translation Engine (MTE) — технология, используемая для генерации машинного перевода. Движки могут быть предобученными (pre-trained или stock) и обучаемыми (custom).

Постредактирование МП (ПРМП) / Machine Translation Post-editing (МТРЕ) — перевод текста с помощью технологии машинного перевода. Цель постредактирования — привести перевод к качеству, идентичному человеческому, используя автоматический машинный перевод.

В отрасли существует разделение на «легкое» и «тяжелое» (light/heavy) постредактирование, где первое подразумевает частичное исправление перевода по смыслу, а второе — приведение машинного перевода к качеству, сопоставимому с качеством человеческого перевода.

В компании «ИнТекст» всегда применяется вариант «тяжелого» постредактирования!

Дистанция редактирования (ДР) / Edit Distance (ED) — это показатель, позволяющий определить, насколько полезным оказался МП. С помощью специальной программы мы измеряем количество изменений, необходимых для преобразования перевода, а также общее количество символов. В результате получаем значение в виде процентов для каждого сегмента. Дистанция редактирования не отражает качество перевода, но отвечает на вопрос, насколько полезным оказался МП.

II. Введение

По мере развития информационных технологий постоянно появляются инструменты для ускорения работы переводчиков: текстовые редакторы, автоматические проверки правописания и правильности перевода, электронные словари, системы памяти переводов. Все эти компьютерные средства имеют общее назначение — предложить переводчику подсказки при работе.

На подходящих текстах современные средства машинного перевода дают качественные подсказки, заметно ускоряя перевод. Кроме того, ряд крупных заказов сейчас поступает только с условием использования предварительно подготовленного машинного перевода. При соблюдении определенных условий умение работать с МП позволяет ускорить работу над традиционными заказами, а также принимать заказы с новыми технологическими цепочками.

Важно отметить, что машинный перевод, как бы хорош он ни был, не заменит работу переводчика. Он подходит не для всех тематик, языковых пар и типов документов и не обеспечивает такое качество, чтобы можно было отказаться от редактирования человеком.

В 2019 году мы протестировали результаты машинного перевода самых разных текстов с помощью нескольких движков. В результате получилось выработать критерии, по которым можно определить целесообразность применения МП для того или иного заказа.

Этот документ описывает технологию машинного перевода и порядок работы с ней. В нем приведены полезные советы по оптимизации процесса работы.

III. Применение МП в работе компании «ИнТекст»

В мире существует более 50 коммерческих движков машинного перевода. Они рассчитаны на разные регионы и работают с разными языковыми парами. Качество и стоимость предлагаемого перевода при этом разительно отличаются. Компания «ИнТекст» приложила много усилий к исследованию и совершенствованию своих технологий. Мы сформировали критерии для применения и выбора ДМП — каждый такой проект мы подготавливаем индивидуально.

В первую очередь мы ориентируемся на **информационную безопасность**. Все заказы, которые мы переводим, конфиденциальны, а значит, провайдеры движков должны гарантировать, что загруженная информация не будет использована кем-либо, кроме представителей «ИнТекст». Бесплатные движки МП такого гарантировать не могут, поэтому в заказах мы используем только платные решения.

Второй критерий для выбора — **качество перевода**, предлагаемое машиной. Чтобы сформировать этот показатель, мы протестировали наиболее распространенные движки для наших языковых пар и тематик. Мы подготовили и проверили широкую выборку тестовых работ. В качестве критериев использовали впечатления переводчиков, а также цифровой показатель дистанции редактирования.

Важным аспектом для качества является то, что выбранные движки работают по принципу нейронного машинного перевода. Эта технология в большинстве случаев позволила улучшить качество машинного перевода и сократить затраты времени переводчика по сравнению со статистическими движками.

Кроме того, важно упомянуть технологию обучения движков машинного перевода. «ИнТекст» владеет этой технологией и выполняет обучение для отдельных проектов и тематик. Благодаря

обучению качество автоматического перевода возрастает — машина понимает и учитывает требования к тексту, стиль, терминологию. В результате перевод выходит более последовательным и лаконичным.

Использование такой технологии требует определенной подготовки ТМ, а также дополнительного бюджета. Поэтому зачастую ее целесообразно применять для больших проектов с постоянным потоком работ.

В результате перевода с помощью МП должен получиться готовый к использованию текст. В последующих разделах мы собрали практические советы, позволяющие достигнуть этого результата более эффективно. Эта информация основана на опыте наших сотрудников, участвовавших в тестировании различных ДМП.

IV. Работа переводчика с использованием МП

Постредактирование как услуга успела укрепиться в отрасли — в открытом доступе есть стандарты переводческих ассоциаций, на конференциях обсуждают нюансы и лучшие способы работы с технологией, не так давно даже был выпущен [стандарт ISO](#).

Важным вопросом остается принадлежность задачи к направлению перевода или редактирования.

В компании «ИнТекст» полагают, что постредактирование больше соответствует переводу. Подготовленный машинный перевод воспринимается как подсказка переводчику, а не как готовый продукт. Кроме того, для приведения качества к уровню перевода человеком необходимо повторно проверить текст после постредактирования.

Можно сказать, что принятый в отрасли термин «постредактирование» фактически означает перевод с использованием подсказок на основе автоматически подготовленных текстов. Поэтому то, что в отрасли называется постредактированием, мы считаем переводом с помощью МП.

Что касается задачи переводчика, который будет использовать машинный перевод, то ее смогли лаконично описать специалисты Национального института стандартов и технологий США:

[Make the MT output have the correct meaning, using understandable English, in as few edits as possible.](#)

Это значит, что при работе потребуется соблюдение трех ключевых условий:

- 1) корректной передачи смысла оригинала;
- 2) использования соответствующего и понятного стиля изложения;
- 3) максимально полного применения машинного перевода, чтобы сократить количество исправлений.

Чтобы соблюсти эти условия, стоит помнить, что:

- часто существует несколько альтернативных способов изменить машинный перевод. Они все могут быть правильными. В такой ситуации лучше ориентироваться на тот, который требуют **меньшего количества исправлений**;

- перевод текста должен всегда передавать **тот же смысл, что и оригинал**, нельзя добавлять или удалять информацию;

- при переводе нужно учитывать **стиль изложения** в целевом языке. Следует избегать подстрочного перевода;

- нужно стараться уменьшить **количество исправлений**, когда это можно сделать без ущерба для смысла или понимания перевода. Каждое вставленное/удаленное/измененное слово или знак препинания считается исправлением.

V. Советы по работе с машинным переводом

1) Непоследовательный перевод

Машина, в отличие от человека, «мыслит» посегментно. Она не способна определять общий контекст всего текста, запоминать слова из предыдущих сегментов и мыслить последовательно.

Это очень заметно, когда в разных сегментах встречается одно и то же слово.

Человек в таком случае сначала проводит исследование и подбирает наиболее подходящий вариант перевода, запоминает его и в дальнейшем использует этот перевод.

Машина же не запоминает, какой вариант перевода она уже выбирала — в каждом сегменте исследование происходит заново. Из-за этого одно и то же слово может быть переведено по-разному.

Чем больше значений у слова, тем сложнее выбор для программы. Например, *listed company* — «компания с биржевым листингом» легко может оказаться «компанией в списке».

Хорошей привычкой при работе с МП станет ведение глоссария.

Благодаря функционалу SDL Trados Studio глоссарий можно создавать и наполнять в процессе работы с текстом, что практически не отнимает времени и позволяет перепроверить себя.

2) Короткие или слишком длинные сегменты

Проблема связана с двумя аспектами работы машинного перевода: отсутствием контекста в целом и сложностью составления связей в сложноподчиненных и сложносочиненных предложениях.

Программа может сильно теряться при переводе коротких сегментов из 1—3 слов. Она рассматривает эти сегменты обособленно, **не понимая общий контекст**. Поэтому при переводе в одних случаях она определит тематику правильно и корректно выберет терминологию, а в других ошибется и выберет совершенно не то.

Документы, состоящие по большей части из такого типа сегментов, не подходят для машинного перевода. Хотя, если движок МП обучен, эту проблему удастся устранить.

В длинных предложениях проблема отличается — программа **путается в связи между предложениями и его частями**. Часто в результате выходит неправильный перевод. Проблема усугубляется, если в исходном тексте проблемы с пунктуацией.

Стоит помнить о рисках, возникающих в слишком коротких или длинных сегментах.

3) Стиль и тон текста

С переходом технологии машинного перевода от статистической к нейронной качество построения предложений, передачи стиля и формулировок значительно улучшилось.

Это дало возможность использовать машинный перевод даже в текстах маркетинговой направленности.

Но при этом важно помнить, что программе все еще свойственны некоторые ошибки.

Существуют **сложности с синтаксисом** — программа может калькировать его на основе исходного текста. Такая особенность будет более всего ощутима при работе с материалами для публикаций.

Еще один типичный прием для программы — использование **однокоренных слов на близком расстоянии** друг от друга. Программа не способна понять, что такой выбор слов вредит стилю повествования.

Также стоит учитывать роль **контекста для стиля** перевода. Программа может перевести текст неоднородно в плане стиля или же, например, не учесть сложившиеся обороты в какой-то определенной отрасли. В текстах про Италию palazzo может быть переведено как «дворец», хотя в части случаев устоялся термин «палаццо». Слово chapel может быть «часовней», хотя применительно к католическим храмам лучше использовать «капелла».

Программа не может различить **формальное и неформальное обращение**. Проблема тона особенно актуальна для переводов с английского языка. Правильность обращений на «вы» или «ты» необходимо проверить вручную, так как полагаться на программу в этом вопросе нельзя.

4) Ошибки исходного текста

Важный аспект, который помогает в работе с МП, — внимательное наблюдение за исходным текстом. Технология пока плохо реагирует на **опечатки в оригинальном тексте**.

Ситуации, которые не вызвали бы никаких сложностей у человека, для программы становятся критическими. Например, словосочетание The folowing matix вместо the following matrix приводит к переводам типа: «утечка банкомата», «складывающийся матикс» и «фоловинг матикс».

Лучше предавать значение подобным случаям сразу во время чтения исходного текста.

5) Работа с тегами

Разные движки машинного перевода работают с тегами по-разному. Наиболее эффективные движки зачастую не имеют каких-то значительных проблем в этом аспекте — теги переносятся в текст перевода удачно. Иногда случаются лишь незначительные недоработки, когда, например, между тегом и словом добавляется пробел.

Если же используется менее продвинутый движок машинного перевода (например, по требованию клиента), есть вероятность столкнуться с более сложными случаями. Мы наблюдали ситуации, когда теги воспринимались программой как разрыв в предложении. Каждая часть предложения до и после тега воспринималась как отдельная.

Мы просим сообщать о таких сложностях сразу при обнаружении!

6) Аномальные ошибки

При работе с машинным переводом иногда можно столкнуться с ошибками, которые выглядят необъяснимо. Например, был случай, когда вариант ответа yes один из движков перевел как «олимпиада».

Мы предполагаем, что такого рода переводы возникли из-за ошибок в материалах, на которых создавался и обучался движок. Это могла быть нарушенная сегментация или же банально неправильный перевод.

Ни один движок не застрахован от такого рода неполадок, и предусмотреть такие случаи нельзя. Поэтому работа с машинным переводом требует повышенной внимательности.

VI. Оценивание качества машинного перевода

Компания «ИнТекст» оценивает все заказы, в которых был применен машинный перевод. Чтобы оценка была полной, учитываются два фактора: личные впечатления переводчика и показатель дистанции редактирования.

Впечатления переводчика

Нам важно получать обратную связь непосредственно от переводчиков. Для этого мы используем веб-формы. В каждой работе с машинным переводом будет особый пункт чек-листа. Это просьба оценить машинный перевод:

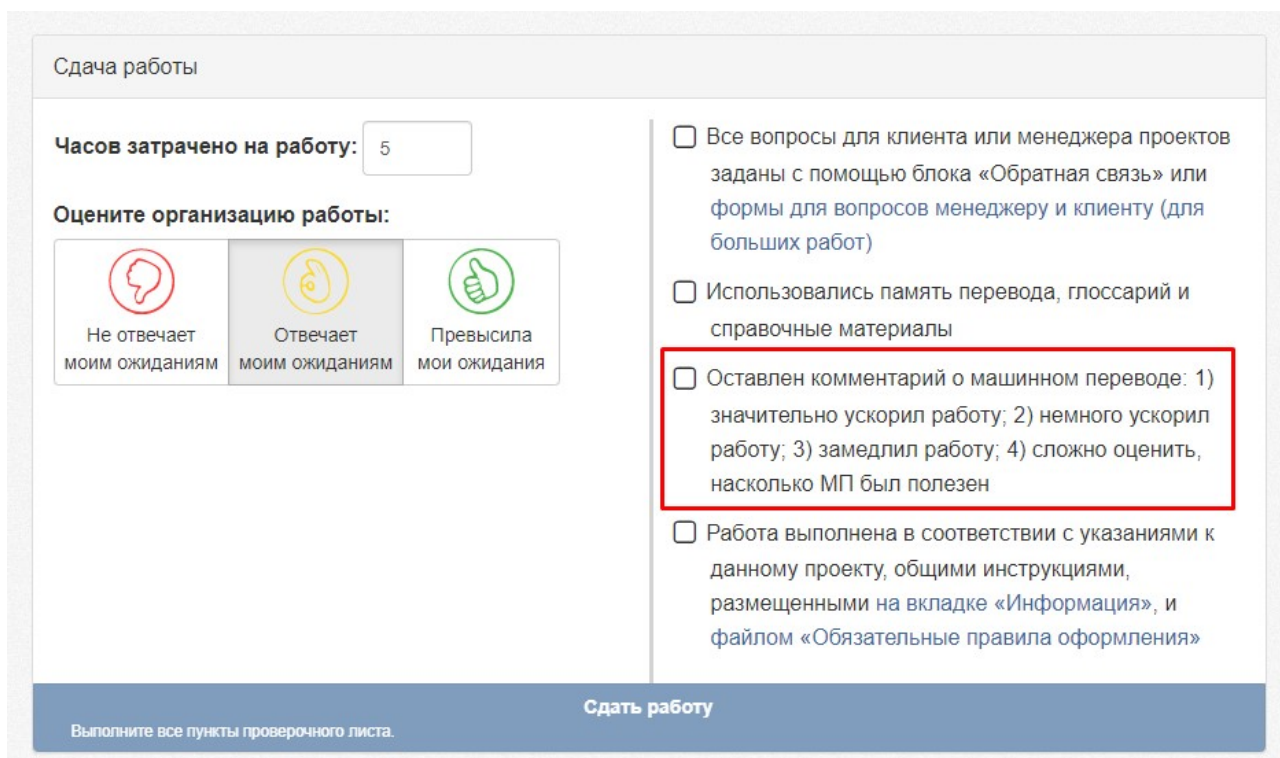


Рисунок 1. Чек-лист для работ с МП

По окончании работы нужно оставить соответствующий комментарий в блоке для обратной связи.

Дистанция редактирования

Это процентный показатель, отражающий степень схожести машинного перевода и измененного текста. Показатель является лишь косвенным отражением качества, его главная цель — дать представление о трудозатратах переводчика при работе с МП. Он означает, сколько нужно было выполнить действий, чтобы изменить текст, и показывает соотношение с общим количеством

символов. Вычисление происходит посегментно, а для всего текста выводится средневзвешенное значение.

Дистанцию редактирования для текста с МП можно трактовать таким образом:

- <50 % — минимальная польза;
- 50—60 % — средняя польза;
- >60 % — высокая польза.

Во время оценивания работ с машинным переводом наш специалист сравнивает впечатления переводчика со средней дистанцией редактирования для сегментов с МП.

Если противоречий нет, он отмечает качество машинного перевода, а если есть — пытается определить причину расхождения. Возможно, для таких работ поможет смена движка машинного перевода или обучение собственного движка. В случаях, когда улучшить качество машинного перевода не выходит, старший менеджер проектной группы обсудит с клиентом целесообразность использования МП в этом проекте.

VII. Нештатные ситуации

1) Качество машинного перевода оказалось недостаточным

Несмотря на то что в компании «ИнТекст» выработаны критерии для заказов, в которых может быть применен машинный перевод, нельзя исключать ситуации, когда машина не даст ожидаемого преимущества.

Мы заинтересованы в том, чтобы переводчики в такой ситуации не ощутили негативных последствий. Если машинный перевод не принес пользы, компания оплатит выполнение работы полностью по обычному тарифу перевода.

Для этого нужно в форме обратной связи на веб-формах сообщить о сложностях и указать, что работу нужно оплатить как обычный перевод.

2) Переводчик предпочитает какой-то конкретный движок машинного перевода

Некоторые переводчики уже имеют свой устоявшийся опыт работы с определенными движками машинного перевода. Они знают их сильные и слабые стороны и понимают, какого результата ожидать в конкретном тексте.

Если существуют такие предпочтения, то о них нужно сообщить на этапе распределения заказа.

Мы постараемся подготовить перевод с нужным движком машинного перевода.

VIII. Условия оплаты

Оплата за перевод с использованием МП в компании «ИнТекст» осуществляется по словам. Тариф составляет 80 % от тарифа за обычный перевод.

Тариф за машинный перевод считается только для новых слов и совпадений 50—74 %. Остальные совпадения обрабатываются без помощи МП и оплачиваются по тарифам обычного перевода.

Тариф не отличается для различных движков машинного перевода. В случае использования обученного движка тариф не снижается.

Компания «ИнТекст» заинтересована в увеличении заработка своих партнеров. Если машинный перевод не ускорил работу над заказом, пожалуйста, сообщите об этом через окно обратной связи на веб-формах. Вы имеете право запросить оплату за заказ по тарифу обычного перевода.

Чем выгоден машинный перевод?

Главное преимущество машинного перевода — ускорение в работе. Согласно нашим тестам, оно может составлять от 20 до 100 % в зависимости от тематики текста, движка машинного перевода и навыков переводчика. А значит, увеличение скорости может быть выше разницы в тарифах постредактирования и обычного перевода. Эффективно также чередовать перевод с помощью МП и обычный.